Leveraging Causal Information for Multivariate Timeseries Anomaly Detection

Lukas Heppel
University of Würzburg, Germany
German Aerospace Center (DLR), Lampoldshausen, Germany

Andreas Gerhardus
German Aerospace Center (DLR), Jena, Germany

Ferdinand Rewicki
German Aerospace Center (DLR), Jena, Germany

Jan Deeken
German Aerospace Center (DLR), Lampoldhausen, Germany

Günther Waxenegger-Wilfing
University of Würzburg, Germany
German Aerospace Center (DLR), Lampoldhausen, Germany

Abstract

Anomaly detection in multivariate timeseries is used in various domains, such as finance, IT, or aerospace, to identify irregular behavior in the used applications. Prior research in anomaly detection has focused on estimating the joint probability of all variables. Then, anomalies are scored based on the probability they receive. Thereby, the variables' dependencies are only considered implicitly. This work follows recent work in anomaly detection that integrates information about the causal relations between the variables in the timeseries into the detection mechanism. The causal mechanisms of the variables are then used to identify anomalies. An observation is identified as anomalous if at least one of the variables it contains deviates from its regular causal mechanism. These regular causal mechanisms are estimated via the conditional distribution of a variable given its causal parent variables, i.e., the variables having a causal influence on a variable. We further develop previous work by gathering information about the causal parents of the variables by applying causal discovery algorithms adapted to the timeseries setting. We apply Conditional Kernel Density Estimation and Conditional Variational Autoencoders to estimate the conditional probabilities. With this causal approach, we outperform methods that rely on the joint probability of the variables in our synthetically generated datasets and the C-MAPPS dataset, which provides simulation data of turbofan engines. Moreover, we investigate the causal approach's inferred scores on the C-MAPPS dataset to gather insights into the measurements responsible for the prediction of anomalies. Furthermore, we investigate the influence of deviations from the true causal graph on the anomaly detection performance using synthetic data.

2012 ACM Subject Classification Computing methodologies → Anomaly detection

Keywords and phrases Anomaly Detection, Causal Discovery, Multivariate Timeseries

Digital Object Identifier 10.4230/OASIcs.DX.2024.11

1 Introduction

Sensors generate data throughout the operational lifespan of assets across various domains, including aerospace and manufacturing [9]. These assets, such as turbofan engines, are susceptible to deterioration and spontaneous faults [26]. Consequently, one of the primary objectives when operating these assets is to assess their condition. Once an asset has reached a certain point of degradation, it is necessary to maintain or replace it to ensure the safety of ongoing operations [10, 9]. The continuous monitoring of assets with the assistance of sensors

35th International Conference on Principles of Diagnosis and Resilient Systems (DX 2024).
Editors: Ingo Pill, Avraham Natan, and Franz Wotawa; Article No. 11; pp. 11:1–11:18
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

generates multivariate timeseries data. In recent years, data-driven approaches have emerged to monitor the health status of these assets, which is coupled with detecting anomalies in multivariate timeseries [36, 9]. In anomaly detection (AD), the objective is to identify data samples that deviate from the prevailing patterns within the dataset [5]. An increase in the number of anomalous observations may indicate asset degradation [9].

Several approaches exist to tackle AD in multivariate timeseries, which consider the multivariate timeseries holistically and only implicitly consider the variables' dependencies. These approaches, for example, rely on Autoencoders (AEs) and variants thereof, such as the Variational Autoencoder (VAE), which estimate the joint probability of individual observations or of observations in a row to describe the regular behavior of the system under consideration [9, 37]. Inference with these methods is then done by assessing the likelihood of given observations. Lower probabilities indicate that an observation could be anomalous [37]. However, when modeling the joint probability of the individual observations directly, we do not explicitly consider the relations of the variables. Recent work on AD in multivariate timeseries explicitly considers the relations among the individual sensors uni-variate timeseries [4, 35]. However, the considered relations are not the causal relations of the underlying system. Thus, recent work by Yang et al. [33] integrates the causal relations amongst the component timeseries into AD. Using the causal relations, the estimation of the joint probability of a sample is factorized into a product of the conditional distributions of the variables given their causal parents, i.e., the variables causing a variable. The conditional distributions are used to characterize the causal mechanisms of the variables. In the causal approach, an anomaly is present if the causal mechanism is disturbed at least on one variable of the multivariate timeseries [33].

However, the causal relations among the variables are often unknown. One approach to obtaining the causes and effects present in a process is to intervene with a system to conduct experiments actively. This is often impracticable since it can be time-consuming, cost-intensive, or requires domain knowledge. However, we can gather multivariate timeseries data from the system through observation [23]. Then, we can reconstruct the timeseries graph embodying the causal parents of the variables based on observational data via Causal Discovery (CD) [24]. The timeseries graph contains the variables at the different timesteps as nodes, and directed edges indicate the causal relations. Thereby, an edge is directed from the causal parent to the child.

In this work, we further develop the causal approach by Yang et al. [33] by applying constraint-based CD algorithms tailored for the multivariate timeseries case, PCMCI [25], and PCMCIplus [22]. Using the reconstructed timeseries graph, we estimate the conditional distribution given its causal parents for every variable. For inference, we use the ensemble of estimates to check whether the causal mechanism of at least one variable is deteriorated. To analyze the effect of the causal relations used, we generate synthetic data that we model as a structural causal process. This way, we know the ground truth causal relations and can evaluate the effects of using relations that differ from the correct ones. Furthermore, we benchmark the causal approach on the publicly available C-MAPPS dataset, which provides simulations of turbofan engines. In addition, we use the inferred anomaly scores from the causal approach to gather insights into the anomalies' emergence.

Our contributions are as follows:

- We develop the causal approach by Yang et al.[33] further by applying CD algorithms tailored for the multivariate timeseries case.
- We use synthetic data to examine the impact of deviations from the true causal graph on the anomaly detection capability.
- We benchmark the causal approach on the publicly available C-MAPPS dataset.

The remainder of this paper is structured as follows: In Section 2, we give an overview of approaches in anomaly detection for timeseries data and introduce the necessary concepts our approach builds upon in Section 3. Afterward, in Section 4, we describe the approach in detail, and in Section 5, we explain our experiments, which we use to evaluate our models. In Section 6, we discuss this work, point out directions for future work, and end this paper in Section 6 with our conclusion.

This paper is the result of work carried out as part of Lukas Heppel's Master's thesis, which was supervised by Günther Waxenegger-Wilfing.

2 Related Work

This section reviews related work on anomaly detection, focusing on detecting anomalies in multivariate timeseries.

Methods for AD comprise distance-based techniques such as the k-nearest Neighbour (kNN) algorithm. In addition, statistical approaches exist to tackle detecting anomalies in timeseries data [30]. Furthermore, Support Vector Machines (SVMs) are trained to classify anomalies in timeseries data [3, 18, 1]. Moreover, anomalies are detected using Isolation Forests (IFs) [11]. Additionally, Yaacob et al. [32] fit an Auto-Regressive Integrated Moving Average (ARIMA) on the timeseries. Afterward, observations are classified based on the error between the prediction of the ARIMA model for the current value and the actual observed value.

In recent years, deep learning methods have been applied to identify anomalies in multivariate timeseries data [5]. Several approaches use Autoencoders (AEs) to detect anomalies [19, 9, 20]. Jakubowski et al. [9] and Reddy et al. [19] train AEs to reconstruct the input data. Anomaly scores are then obtained as the error between the input and its reconstruction. Other AE variants, besides the vanilla AE, are also used to detect anomalies. For example, Variational Autoencoders (VAEs) [9, 31] allow obtaining the reconstruction probability of the data points and then computing anomaly scores. Zong et al. [37] introduce Deep Autoencoding Gaussian Mixture Model (DAGMM) for AD in multivariate timeseries. This approach consists of an AE, which provides lower-dimensional representations of the input, i.e., the latent representation. The latent representation and the corresponding reconstruction error are then used to estimate the density of a sample using the estimation network, which is trained as a Gaussian Mixture Model (GMM). In addition to the aforementioned AE variants, where the encoder and decoder consist of Feedforward Neural Networks, the approach of Park et al. [14] combines LSTMs with VAEs to encode the temporal dependencies of the timeseries. However, these methods implicitly consider the dependencies among the variables and ignore explicit knowledge, which is sometimes available. In addition, these approaches cannot trivially be used to explain an anomaly since they only provide one score/classification per observation.

Further approaches arise that strive to include knowledge of the dependencies of the variables contained in the multivariate timeseries [4, 6]. Dai et al. [4] use normalizing flows to estimate the joint probability of all variables in the multivariate timeseries. They incorporate a Directed Acyclic Graph (DAG) into the probability estimation to factorize the joint probability into a product of conditional densities. The DAG used for factorizing the joint probability is optimized with the normalizing flow. In addition, Deng et al. [6] apply graph attention-based forecasting of future sensor values for AD while relying on graph structures that represent as nodes the sensors, which are connected by edges with the k nodes they have the most similar timeseries with. However, the variables' relations used in these approaches are not necessarily causal relations.

Therefore, Yang et al. [33] explicitly consider the relations among the variables. Thus, they reconstruct the causal relations present in the multivariate time via the CD algorithm PC and score-based CD algorithms. Based on the causal relations, they estimate the conditional distribution of each variable, given its causal parents, via CVAEs. For the inference, they predict an anomaly if one variable's model predicts an anomaly smaller than a given threshold.

Our approach differs from the methodology proposed by Yang et al. [33] in detecting the causal relations and the used density estimation techniques. To this end, we rely on the constraint-based approaches PCMCI [25] and PCMCIplus [22], which are adaptions for the timeseries setting and show superior performance compared to the PC algorithm [28], as shown by Runge et al. [25, 22]. Moreover, we examine in greater detail the impact of deviations from the true graph structure on the AD performance of the causal approach.

3 Preliminaries

This section introduces the necessary concepts this paper builds upon.

3.1 Reconstruction-based AD

This work is concerned with detecting anomalies in the multivariate timeseries setting. Thus, we define a multivariate timeseries as $x = (x^1, \dots x^d) \in \mathbb{R}^{T \times d}$, which is an ordered sequence of d univariate timeseries $x^i \in \mathbb{R}^T$, where each describes a feature x^i over a period of length T. We denote the value of feature i at time t by $x^i_t \in \mathbb{R}$. Furthermore, $x_t = (x^1_t, \dots, x^d_t) \in \mathbb{R}^d$ is the vector describing the value of all features at time t.

Autoencoders (AEs) are used to identify anomalies [9]. The AE architecture consists of an encoder and a decoder. The encoder maps an input vector $x \in \mathbb{R}^{d_{input}}$ into a latent representation $z \in \mathbb{R}^{d_{latent}}$. In most cases, AEs apply a bottleneck, i.e., the dimension $d_{latent} < d_{input}$ of the latent space is smaller than the dimension of the input. Introducing a bottleneck pushes the AE to learn meaningful latent representations from the input. Next, the decoder transforms the latent representation z into a reconstruction $\hat{x} \in \mathbb{R}^{d_{input}}$ of the input dimension. We train the AE to minimize the error between the input x and the reconstruction $\hat{x}$ using the Mean Squared Error (MSE). For inference, the reconstruction error of a sample can be used as an anomaly score, assuming that anomalous data differs from the known training data and, therefore, cannot be properly reconstructed.

The Variational Autoencoder (VAE) is a variant of the AE that estimates the distribution of a given dataset [13]. Thereby, the assumption is made that the generative process of the data points includes continuous latent variables z. To this end, the VAE's encoder approximates the distribution $q(z \mid x)$, and the decoder estimates the distribution $p(x \mid z)$. We can then formulate the variational lower bound of $log\, p(x_i)$, which is also referred to as evidence lower bound (ELBO), as follows [13]:

$$log\, p(x_i) \geq \mathbb{E}_{z \sim q(z|x_i)}[log\, p(x_i \mid z)] - D_{KL}(q(z \mid x_i) \mid\mid p(z)) \tag{1}$$

Thereby, D_{KL} represents the Kullback-Leibler (KL) divergence [34]. The initial term represents the KL divergence between the encoder distribution $q(z \mid x_i)$ and the prior $p(z)$, which we assume to be a standard normal distribution $\mathcal{N}(0, 1)$. Consequently, the learned posterior distribution is regularized to be close to a standard normal distribution. The second term is the expected negative log-likelihood of the input given the latent representation. We compute the expectation by sampling a number of latent representations z_i for which the decoder network provides the reconstructions. These reconstructions are then used to derive the parameters of the decoder distribution, which we assume to be Gaussian. The VAE can then be trained by minimizing the negative ELBO.

To detect anomalies, we are then required to infer the likelihood of a sample that should be assessed whether it's anomalous. We assume that anomalies obtain smaller likelihoods since they deviate from the distribution of the normal samples. Throughout this work, we measure a sample x_i's density $p(x_i)$ by the decoder's distribution based on the reconstructions resulting from the latent samples $z_i \sim p(z \mid x_i)$.

3.2 CVAE

The Conditional Variational Autoencoder (CVAE) [27] is an extension of the VAE that allows for conditional inputs. The CVAE allows us to guide its generative process via attributes, e.g., in image generation [2]. Thus, the CVAE estimates the conditional log-likelihood of the data. Consequently, the encoder is extended to approximate $q(z \mid x, c)$. Moreover, the decoder is extended to model $p(x \mid z, c)$. Then, to estimate the conditional log-likelihood, the variational lower bound of a data point x_i given c_i is defined as follows [27]:

$$log\ p(x_i \mid c_i) \geq \mathbb{E}_{z \sim q(z|x_i,c_i)}[log\ p(x_i \mid z, c_i))] - D_{KL}(q(z \mid x_i, c_i) \mid\mid p(z \mid c)) \tag{2}$$

We assume the prior $p(z \mid c) = \mathcal{N}(0, 1)$ to be a standard Normal distribution. The CVAE can then be trained by minimizing the negative variational lower bound.

3.3 Causal Discovery

The AD approach proposed in this work requires information about the causal relations underlying the observed system, which is the origin of the multivariate timeseries data. However, the causal relations are often unknown and need to be discovered. We could intervene and actively conduct experiments to gather information about the causal relationships in the system under consideration. However, active experimentation is often impossible because it is time-consuming, cost-intensive, demands expert knowledge, or is infeasible, e.g., in our case, where we do not have the knowledge to work with the simulation software providing our dataset. Nevertheless, the systems under consideration can often be observed to obtain multivariate timeseries data [23].

In CD in the timeseries setting, we aim to reconstruct the underlying timeseries graph that embodies the temporal dependency structure of a given dynamic system. For this purpose, we assume a *discrete-time structural causal process* $x_t = (x_t^1, \dots x_t^d)$, which is defined as follows [22]:

$$x_t^i = f_i(pa(x_t^i), \eta_t^i), \tag{3}$$

with f_i denoting arbitrary measurable functions that non-trivially depend on their arguments. Moreover, $pa(x_t^i)$ are the causal parents of the variable x_t^i, and η_t^i represents noise, which is assumed to be independent between the variables and the timesteps.

The *timeseries graph* $G = (V, E)$ is defined as follows:

- The set of nodes V represents the variables x_t^i at the different timesteps t.
- E is the set of edges containing an edge $x_{t-\tau}^j \to x_t^i$ iff $x_{t-\tau}^j \in pa(x_t^i)$

Furthermore, we define the variables a variable x_t^i causally depends on, as its causal parents $pa(x_t^i) \subset (x_t, x_{t-1}, \dots)$. Thereby, we refer to links with $\tau = 0$ as contemporaneous links and those with $\tau > 0$ as lagged links. In the case $i = j$, we refer to such links as autodependency. In addition, we assume the graph to be acyclic. Moreover, we assume the case of Causal Stationarity, which means that the causal relationships are assumed to be invariant in time, i.e., if there exists a link $x_{t-\tau}^j \to x_t^i$ for a t, then we assume $x_{t'-\tau}^j \to x_{t'}^i$ for all $t' \neq t$.

Alternatively, different parts of a time series can be represented by different graphs to model non-stationarity [24]. Theoretically, the graph is infinite. However, we consider the graph only up to a maximum time lag τ_{max}.

The predominant methodologies employed in the field of CD can be broadly classified into three main categories: score-based, asymmetry-based, and constraint-based [24]. We are concerned with applying constraint-based methods to detect causal relations in multivariate timeseries. Score-based and asymmetry-based approaches are still in their infancy for timeseries [24]. One popular approach for CD is the PC algorithm [28], originally developed for independent and identically distributed data.

The PC algorithm is the base for the PCMCI [25], and the PCMCIplus [22] algorithms, which we use for multivariate timeseries data in this paper. Constraint-based methods for CD, such as the PC algorithm, are conducted in two steps. In the initial phase, the graph's skeleton is reconstructed. This skeleton is an undirected graph that merely indicates whether two nodes are connected. In this phase, we employ an iterative approach to ascertain which pairs of variables are conditionally independent based on tests of the form:

$$H_0 : x^i \perp\!\!\!\perp x^j \mid x^k \qquad\qquad H_1 : x^i \not\perp\!\!\!\perp x^j \mid x^k,$$

thereby H_0 states the conditional independence of x^i and x^j given x^k, whereas H_1 states their conditional dependence. In the second phase, the skeleton's links are oriented, i.e., the directions for the undirected edges are determined. Therefore, time provides an orientation for lagged links, as causes precede effects. The orientation of contemporaneous links ($\tau = 0$) is not a straightforward process and is determined by applying additional rules. In general, constraint-based approaches can only detect the contemporaneous graph's Markov equivalence class [23]. The Markov equivalence class is the set containing all DAGs that embody the same conditional independence [16]. Thereby, the directions of the edges must not be the same in all DAGs of the Markov equivalence class. Thus, links between variables x_t^i and x_t^j where both directions, i.e., $x_t^i \to x_t^j$ and $x_t^i \leftarrow x_t^j$ are present in at least one DAG of the equivalence class, are represented as unoriented links in the resulting graph.

Since we can only measure statistical dependencies from data in the constraint-based approaches, we need several assumptions to interpret the graph structure obtained from constraint-based approaches as the timeseries graph, which describes the causal relations. We make the following assumptions to be able to apply the later discussed CD algorithms:

- *Causal Sufficiency*: This assumption states that no other unobserved variables directly or indirectly influence any other set of observed variables [25, 29].
- The *Causal Markov condition* states that the connectivity from the causal structure imprints the marginal and conditional (in)dependencies into the observed distribution [24, 29].
- The *Faithfulness* assumption states that all conditional independencies result from the causal structure [24, 29].

Using the Causal Markov condition and the Faithfulness condition, the following equivalence holds [24]: $x^i \perp\!\!\!\perp x^j \mid x^k \iff x^i$ d-sep $x^j \mid x^k$. This equivalence relates the graph structure, i.e., the contained d-separations [24], to the conditional independencies detected in the data.

The PCMCI algorithm addresses issues of the conditional independence framework when applied to the timeseries setting, e.g., the PC algorithm is prone to high false positive rates when facing auto-correlation [25]. In addition to the aforementioned assumptions, the PCMCI algorithm relies on the assumption that no contemporaneous links exist ($\tau = 0$) since PCMCI cannot orientate these. However, when we cannot assume the absence of contemporaneous links, we apply the PCMCIplus algorithm to detect contemporaneous graph structures up to the Markov equivalence class [22]. It is possible that some links remain unoriented, which indicates that both directions exist for this link in the Markov equivalence class. In addition, conflicts may arise when applying the orientation rules, in which case the links are marked accordingly. One advantage of constraint-based approaches for CD is that they can be combined with several variants of conditional independence tests based on the dependencies assumed in the observed data. Thereby, they test $x^i \perp\!\!\!\perp x^j \mid x^k$. In the following, we provide information on the conditional independence tests considered in this paper:

- The Partial correlation test [25] (ParCorr) assumes a linear additive noise model with Gaussian noise. The test consists of two steps. First, ordinary least square regressions for x^i and x^j on x^k are fitted. Afterward, the Pearson correlation test is applied to the residuals.
- The Robust Partial Correlation [25] (RobParCorr) assumes the existence of a linear additive noise model with Gaussian noise and that the observed variables emerge thereof by component-wise transformation. RobParCorr can be applied when linear dependencies exist among the variables. In addition, RobParCorr is suited for different marginal distributions. Thus, RobParCorr extends ParCorr by first transforming the marginal distributions of the variables to standard normal marginals. Afterward, ParCorr is applied.

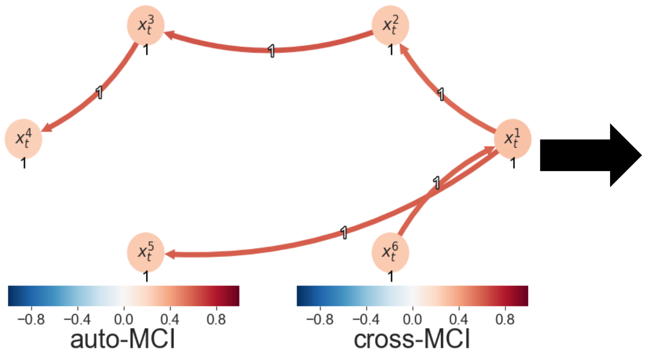
When applying the CD algorithms, we must determine the appropriate conditional independence test. To ascertain the dependency types among the variables, we visually inspect the scatter plots of pairs of variables with varying time lags. Furthermore, we visually inspect the marginal densities of the variables and the joint densities of the pairs of variables. Based on these observations, the most appropriate test for using the CD algorithm can be selected. Furthermore, we need to determine the minimum time lag τ_{min} and the maximum time lag τ_{max}. The choice of τ_{min} has to be made based on assumptions on the data. Thereby, the case that it is necessary to set $\tau_{min} = 0$ can occur when the sampling rate of the data is smaller than the rate at which the effects occur in reality. Domain knowledge is thereby a good help. The selection of τ_{max} is contingent upon the maximum anticipated time lag within the data. Consequently, the optimal value for τ_{max} can be determined by applying domain knowledge or by investigating the lagged dependencies of the pairs of variables, which refers to the lagged cross-correlation function between the pairs of variables, i.e., by testing the unconditional independence of the pairs of variables for varying time lags [25].

4 Methodology

This section describes the methodology we evaluate in this paper to detect anomalies in multivariate timeseries. First, we describe the decomposition of the task by incorporating the causal relations among the variables as described by Yang et al. in [33]. We explain how to obtain causal relations by conducting CD. Then, we outline the model fitting process for estimating the necessary (conditional) distributions. Finally, we present the inference of anomaly scores from the estimated models.

4.1 Factorization using Causal Relations

We assume $x = (x^1, \dots, x^d) \in \mathbb{R}^{T \times d}$ to be a multivariate timeseries as defined in Section 3. Furthermore, we assume x to represent the regular behavior of the considered system, i.e., not containing anomalous points. Then, the task is to detect anomalous observations occurring



$$
\begin{aligned}
M_1 &\approx \mathrm{P}(x_t^1 | x_{t-1}^1, x_{t-1}^6)\\
M_2 &\approx \mathrm{P}(x_t^2 | x_{t-1}^2, x_{t-1}^1)\\
M_3 &\approx \mathrm{P}(x_t^3 | x_{t-1}^3, x_{t-1}^2)\\
M_4 &\approx \mathrm{P}(x_t^4 | x_{t-1}^4, x_{t-1}^3)\\
M_5 &\approx \mathrm{P}(x_t^5 | x_{t-1}^5, x_{t-1}^1)\\
M_6 &\approx \mathrm{P}(x_t^6 | x_{t-1}^6)
\end{aligned}
$$

Figure 1 Illustration of how we obtain the models estimating the conditional distributions representing the causal mechanism of the variables from the process graph, which summarizes a timeseries graph resulting from a CD. The links indicate the causal parent relations and the time lag for the corresponding relation is noted at the link. For the autodependencies, the time lag is depicted next to a node. The color of the links and the nodes indicate the value of the conditional independence tests.

after time T. To detect anomalies x_t with $t > T$, we aim to identify data points x_t, where at least one $x_t^i \in x_t$ deviates from the causal mechanism so that the observed x_t^i and its causal parents $pa(x_t^i)$ deviate from the conditional distribution $P(x_t^i \mid pa(x_t^i))$, which characterizes the ordinary causal relation between x_t^i and its causal parents $pa(x_t^i)$. The causal relations of the timeseries can be represented by the corresponding discrete-time structural causal process $x_t^i = f_i(pa(x_t^i), \eta_t^i)$, as described in Section 3. Using the causal relationships between the variables, the joint distribution can be computed according to the Markov factorization as follows [16, 33]:

$$P(x_t) = \prod_{x_t^i \in x_t} P(x_t^i \mid pa(x_t^i)) \tag{4}$$

Thus, we decompose the task of estimating the joint distribution $P(x_t)$ into estimating one model M_i for the conditional distribution $P(x_t^i \mid pa(x_t^i))$ per variable $x_t^i \in x_t$ given its causal parents $pa(x_t^i)$. Figure 1 illustrates how we obtain the models M_i from an exemplary process graph, which summarizes a timeseries graph.

In some cases, we know these relations and can directly use them. In general, we do not know these relations. Therefore, we are concerned with retrieving the causal relations amongst the variables $x_t^i \in x_t$. To this end, we apply CD to obtain the causal relations. We execute a data analysis of the data to be discovered. Based on the dependency types of the variables, their joint and marginal distributions, their τ_{min}, and τ_{max}, and their lagged cross-correlations, we can select the adequate CD algorithm in combination with the proper conditional independence test. The result of the CD is then the timeseries graph representing the causal dependencies among the variables $x_t^i \in x_t$. However, the result of the PCMCIplus algorithm potentially provides unoriented links since the algorithm can only detect the Markov Equivalence class of contemporaneous links. Furthermore, we can encounter conflicting links where the application of the orientation rules is impossible. Thus, we consider the following post-processing options for the CD results to obtain the causal parents for each variable:

- One approach only considers the directed links in the detected timeseries graph. The unoriented and the conflicting links are not further considered. We refer to this technique as partial factorization throughout this paper.

- A further approach would be to consider using both possible directions of an unoriented link $x_t^i - x_t^j$, i.e., $x_t^i \rightarrow x_t^j$ and $x_t^i \leftarrow x_t^j$. The conflicting links are also not considered further.

4.2 Inference of Anomaly Scores

After discovering the causal relationship of the variables, we train one model M_i for every variable $x_t^i \in x_t$ to estimate the corresponding conditional probability distribution $P(x_t^i \mid pa(x_t^i))$. We follow Yang et al. [33] and employ a CVAE to estimate the respective density functions. Additionally, we consider Conditional Kernel Density Estimation (CKDE) [17]. In theory, further approaches can be integrated to estimate the conditional distributions. We estimate the conditional distributions $P(x_t^i \mid pa(x_t^i))$ using normal data samples $\{(x_t^i, c_t^i) \mid 1 \leq i \leq T\}$, with $c_t^i = pa(x_t^i)$ and T being the length of our training data.

Depending on the obtained timeseries graph, variables $x_t^i \in x_t$ may exist without causal parents $pa(x_t^i) = \varnothing$. This paper evaluates several methods to estimate these variables' distribution $P(x_t^i)$, i.e., the respective density functions. The following approaches are considered:

- We estimate the density $p(x_t^i)$ using a VAE. However, we encounter that variables without causal parents lead to one-dimensional inputs, so we cannot apply dimensionality reduction in the VAE.
- A further approach is to estimate $p(x_t^i)$ for a variable x_t^i without causal parents using Kernel Density Estimation.

To estimate $p(x_t^i)$, the corresponding training data is defined as follows $\{x_t^i, \mid 1 \leq i \leq T\}$. In addition, if multiple variables x_t^i exist without causal parents, we can apply a VAE to estimate the joint probability of these variables. We train the respective CVAEs and the VAEs by minimizing the negative variational lower bounds, which are described in Section 3.

After training the models M_i to estimate (conditional) distributions on the training data, we use the ensemble of the models to infer anomaly scores for previously unseen data samples x_t, with $t > T$.

In the following, we describe the inference of anomaly scores from the ensemble of models M_i. For every variable x_t^i, we infer $log\ p(x_t^i \mid pa(x_t^i))$, i.e., the logarithm of the likelihood of the variable's current value x_t^i given the current observed values of its parents. For variables without causal parents, we infer $log\ p(x_t^i)$, the log-likelihood of the current value. We assume that the lower the detected likelihood, the more anomalous a variable's current value is considered to be. Thus, to obtain an anomaly score, where higher values indicate a higher likelihood of an anomaly, we use the negative log-likelihoods as the anomaly scores of the variables. Thus, the anomaly score s_t^i of model M_i for an observation x_t^i is defined as follows:

$$s_t^i = M_i(x_t^i) = \begin{cases} -log\ p(x_t^i) & \text{if } \ pa(x_t^i) = \varnothing, \\ -log\ p(x_t^i \mid pa(x_t^i)) & \end{cases}$$

To obtain the final anomaly score s_t of the ensemble, we consider using the maximum of the scores from all models. Since we assume that one variable violating its causal mechanism is sufficient to mark the complete observation as an anomaly, we consider taking the maximum of all local anomaly scores as the final score:

$$s_t = \max \left\{ s_t^i \mid x_t^i \in x_t \right\} \tag{5}$$

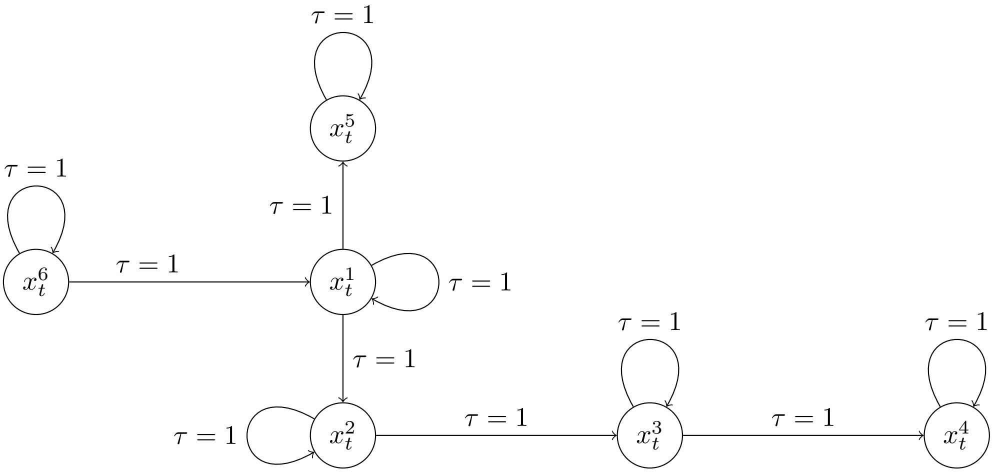


Figure 2 The ground truth process graph defining causal relations in the structural causal process to generate synthetic multivariate timeseries data. The coefficients for the causal dependencies are not depicted. E.g., $pa(x_t^2) = \{x_{t-1}^1, x_{t-1}^2\}$.

5 Experiments

This section presents the experiments conducted to evaluate the proposed methodology. To this end, the methodology is evaluated on synthetically generated data with a known ground truth graph structure. Furthermore, the impact of the applied graph structure on AD performance is examined. Finally, the approach is benchmarked on the publicly available C-MAPPS dataset [26].

5.1 Synthetic Data

We evaluate the causal approach on synthetically generated data. Thus, we define a structural causal process as a generalized additive model. The model is defined as follows [22]:

$$x_t^i = \eta_t^i + a^i x_{t-1}^i + \sum_{x_{t-\tau}^j \in pa(x_t^i) \setminus x_{t-1}^i} c_\tau^j f_\tau^j(x_{t-\tau}^j) \tag{6}$$

with variables $i \in \{1, \ldots, d\}$ and $\eta_t^i \sim \mathcal{N}(\mu_i, \sigma_i)$ denoting the additive Gaussian noise term of variable i at timestep t, which is sampled from a Gaussian distribution with mean μ_i and standard deviation σ_i. In addition, a^i represents the coefficient for the autodependency of variable i, and c_τ^j refers to the dependency coefficient of variable j over time lag τ. Moreover, $pa(x_t^i)$ denotes the set of causal parents $x_{t-\tau}^j$ of the variable i with the corresponding time lag τ and $f_\tau^j(x_{t-\tau}^j)$ corresponds to the dependency function.

Next, we define the ground truth causal relations of the model, i.e., for every variable x_t^i, its causal parents $pa(x_t^i)$. The process graph depicting the causal relations used for our datasets is depicted in Figure 2.

Furthermore, we use for all (auto)dependency coefficients 0.3, and we sample all variables' noise terms from $\mathcal{N}(0, 0.1)$. The applied dependency function is defined as $f(x) = (1 - 4e^{\frac{-x^3}{2}})x$. We use these parameters to generate the multivariate timeseries data to train the models based on regular data.

Moreover, we add anomalous datapoints into the test data. Therefore, we follow Yang et al. [33] and differentiate between the following types of anomalies:

- Intervention: We insert an anomaly at a variable i at time t. The anomaly is then propagated along the causal mechanism to the causal children variables j. In our case, we chose variable x^2.
- Measurement: We insert an anomaly at a variable i at time t. The anomaly is not propagated along the causal mechanism. We imitate a measurement error, i.e., the system works correctly, but, e.g., a sensor reported a wrong value. In our case, we chose variable x^6.
- Effect: We insert an anomaly at a variable i at time t. We chose a variable i, which does not have causal children. In addition, we remove the autodependency of this variable. In our case, we chose variable x^5.

Furthermore, we control the number of anomalies we insert into the test data. Thus, we randomly corrupt ten % of data points as anomalies. Thus, we transform normal data points into anomalous data points by defining a Uniform distribution $\mathcal{U}(2, 10)$, which we use to sample factors to scale a normal sample to be anomalous. We limit the resulting values of anomalous data to the range of the train data. We generate 7000 samples for model training, where we use 15 % for validation. In addition, we generate 3000 samples for testing.

For the causal approach, we apply Conditional Kernel Density Estimation (CKDE) or CVAEs to fit the conditional density of a variable x_i given the causal parents $pa(x_i)$.

Throughout this work, we assume the AEs (and variants thereof) to be symmetric, i.e., the encoder and decoder contain the same number of layers with the same number of neurons but reverse in their order. Furthermore, we assume all variants discussed and used in this paper to apply steady compression, that is, the number of neurons contained in one layer decreases/increases to the next layer by a constant factor. The constant is obtained when knowing the input and latent dimensions, i.e., how much compression should be applied. Doing this removes the need to tune the hidden layer dimensions as additional hyperparameters [9].

For the CVAE, we apply one hidden layer in the encoder and the decoder and a latent dimension of one. We train the CVAE over 50 epochs and apply early stopping with a patience of ten epochs. Furthermore, we sample 64 latent representations per data point. We use the ADAM optimizer [12] for any model training in this work. In the synthetic data experiments we apply a learning rate of 0.001.

We use the implementation by Rothfuss et al. [21] with the default settings. Inference is done as described for the CVAE. In addition, we provide the VAE and KDE as a baseline, which estimates the joint density $p(x_t)$ for all sensors using the train samples x_t. The anomaly score of these models for a sample x_t is then the negative log-likelihood $-log\ p(x_t)$.

For the VAE, we apply two hidden layers in the encoder and the decoder and a latent dimension of two. Furthermore, we sample 64 latent representations per data point. We train the VAE over 100 epochs and apply early stopping with a patience of 10 epochs.

We apply the KDE implementation from scikit-learn [15] with the Epanechnikov kernel and Scotts's rule for the bandwidth selection, which worked best in preliminary experiments. The other parameters are kept default.

Moreover, we report as a benchmark the performance of the conditional density estimate of the variable x_t^i, where the anomalies are injected, given its causal parents $pa(x_t^i)$. We use the negative conditional log-likelihood $-log\ p(x_t^i \mid pa(x_t^i))$ of a variable x_t^i at time t as the anomaly score for inference.

We investigate the impact of the used causal relations. Thus, we apply the causal approach with causal structures that deviate from the ground truth. In the following, we list the considered graph structures:

- We apply the ground truth graph structure.
- We evaluate the fully connected graph for $\tau = 1$, including the autodependencies. This scenario includes, besides all true links, all possible false positives.
- We investigate the graph structure obtained when we apply the PCMCI algorithm to CD in combination with the ParCorr conditional independence test on non-linear data, with $\tau_{min} = 1$, and $\tau_{max} = 2$. This graph structure yields all true links and several false positives (FPs). However, we consider more realistic false positives in this experiment, resulting in lower dimensions for the distributions to estimate than when applying the fully connected graph structure.
- We examine the graph structure consisting of all links at $\tau = 1$ except for the true links. Thus, the used graph structure contains all false negative (FN) and FP links.
- We investigate the graph structure containing no links at all, i.e., we assume no variable is the cause or effect of another.

We report the average precision (AP) in Table 1, and the TPR@{0.05,0.1} in Table 2. The causal approach using CVAEs performs superiorly over the VAE and KDE baselines regarding the AP for all anomaly types. Similar observations can be made regarding the reported TPRs. Furthermore, the individual CVAE of the anomalous variable shows the best AD performance regarding all anomaly types and all reported metrics. Regarding the varying graph structures, we observe the best performance when using the ParCorr graph, which performs better for measurement anomalies than the true graph structure and is on par for intervention anomalies regarding the AP. This indicates the importance of detecting the true causal parents, but also the ability to deal with several false positives (ParCorr graph). However, when the amount of false positives increases, the performance decreases, as can be seen by the performance of the fully connected graph. In addition, when the causal parents are not integrated, as when using no links at all, we obtain performance decreases. Moreover, the true graph structure performs better than all other graph structures in all metrics for effect anomalies and regarding TPR@{0.05,0.1} for measurement and intervention anomalies. We see smaller performance degradations for the differing graph structures for measurement anomalies than for intervention and effect anomalies.

5.2 C-MAPPS

We evaluate the causal approach on the publicly available C-MAPPS dataset [26], which was initially used to predict turbofan engines' Remaining Useful Life (RUL) [8, 26]. The dataset provides multivariate timeseries data from turbofan engine simulations conducted with the Commercial Modular AeroPropulsion System Simulation (C-MAPPS) software [7]. The simulation software can simulate different operation conditions, such as the altitude or the temperature. The dataset comprises four sub-datasets: FD001, FD002, FD003, and FD004. These sub-datasets differ in the simulated operation conditions and the reasons for engine failure. They consist of multiple engine simulations, which can be considered a fleet of engines. Therefore, every dataset contains multiple timeseries, each describing the simulation of one engine. The multivariate timeseries contain 21 measurements/variables of the turbofan engine, such as the fan speed or the temperature at the low-pressure turbine (LPC), for all sensor values, refer to the paper by Saxena et al. [26]. Thus, for every engine, multiple flights are simulated, and the corresponding timeseries reports one measurement value per sensor per flight. The initial wear differs among the engines. We follow Jakubowski

Table 1 Results for AD on the synthetic data with non-linear dependencies. Metric is AP. The best and the second-best results are denoted in bold.

Model	Measurement	Effect	Intervention
CVAE of Anomalous Variable	**0.716**	**0.667**	**0.495**
VAE with x_t	0.434	0.241	0.191
KDE	0.338	0.250	0.215
Causal + CVAE + True Graph	0.621	**0.464**	**0.245**
Causal + CKDE + True Graph	0.470	0.429	0.124
Causal + CVAE + Fully Connected Graph	0.273	0.163	0.190
Causal + CVAE + ParCorr Graph	**0.633**	0.358	0.243
Causal + CKDE + All FPs & FNs	0.604	0.104	0.110
Causal + KDE + No Links	0.514	0.226	0.156

et al. [9] and use the dataset FD004, which they consider the most challenging scenario since six different operation conditions are present and two possible failure modes occur. Training and test sets are provided for every sub-dataset. The train set contains run-to-failure of individual engine simulations. In contrast, the test set's timeseries stop at a point in time before the engines fail. The test set also contains the RUL values of each timeseries, which will be predicted in the original task.

Jakubowski et al. [9] use the dataset for anomaly detection. We follow their definition for labeling the data points as anomalies or normal. Concrete data points with an RUL greater than 130 are assigned to the normal class. Samples with an RUL smaller than 20 are assigned to the anomaly class. The intermediate samples are not considered further. Following the labeling procedure, the test set for RUL prediction yields only a small amount of anomalous samples due to the truncated runs, which often miss the anomalies. Therefore, we use the run-to-failure simulations from the initial train set and assign each run from the C-MAPPS train dataset to the train, the validation, or the test set. In contrast to Jakubowski et al. [9], we do not use labeled validation data to fit a threshold. In Table 3, we depict the number of samples per split. The train and validation splits do not contain anomalous samples since we train unsupervised.

For the causal approach, we first discover the causal relations of the C-MAPPS dataset. Thus, we inspect the joint and marginal densities for the pairs of variables, encountering non-Gaussian densities. In addition, we examine the scatter plots for the pairs of variables, we encounter a strong tendency for linear dependencies. Therefore, we select the RobParCorr test for conditional independence. The analysis of the lagged dependencies, i.e., the unconditional dependencies of the pairs of variables over the various time lags, indicates that the selection of $\tau_{min} = 0$ is adequate, especially since we work with data with a small sample rate, i.e., one measurement per flight. Thus, we apply the PCMCIplus algorithm to be able to obtain

Table 2 Results for AD on the synthetic data with non-linear dependencies. Metric is TPR@FPRs. The best and the second-best results are denoted in bold.

	Measurement		Effect		Intervention	
Model	TPR@0.05	TPR@0.1	TPR@0.05	TPR@0.1	TPR@0.05	TPR@0.1
CVAE of Anomal. Variable	**0.668**	**0.702**	**0.657**	**0.733**	**0.505**	**0.672**
VAE with x_t	0.395	0.565	0.140	0.264	0.148	0.236
KDE	0.313	0.490	0.195	0.346	0.154	0.262
Causal + CVAE + True Graph	**0.626**	**0.678**	**0.458**	**0.581**	0.204	**0.383**
Causal + CKDE + True Graph	0.599	0.643	0.384	0.486	0.079	0.131
Causal + CVAE + Fully Connect. Graph	0.260	0.359	0.110	0.198	0.16	0.242
Causal + CVAE + ParCorr Graph	0.594	0.648	0.359	0.559	**0.233**	0.247
Causal + CKDE + All FPs & FNs	0.585	0.629	0.051	0.116	0.059	0.115
Causal + KDE + No Links	0.548	0.602	0.058	0.385	0.059	0.134

Table 3 The number of samples contained in the datasplits of the C-MAPPS dataset.

	# normal samples	# anomal samples
Train Set	19057	-
Validation Set	5052	-
Test Set	4898	760

also directed contemporaneous links. We select $\tau_{max} = 3$ to integrate the peak in the lagged dependencies at $\tau = 0$ and provide a buffer. We use the implementation from the tigramite package by Runge et al. [22]. Thereby, we do not specify a pc_α value, which is used for the selection of the conditions used in the conditional independence tests, refer to Runge et al. [25, 22]. The remaining parameters of the algorithm are kept default.

Next, we apply the detected graph structure in the causal approach by only using the detected directed links (partial factorization). Thus, we encounter at least one causal parent for every variable. We provide the results for the causal approach based on CVAEs and CKDEs. In addition, we provide a mixture of CVAEs and CKDEs. Thus, we use a CKDE for variables with less than three causal parents and a CVAE for others. For the CVAEs, we keep the hyperparameters as for the synthetic data, except that we evaluate per variable the number of hidden layers, either one or two, and the latent dimension, either one or half of the input dimension of the CVAE (variables dimension plus number of causal parents).

Besides the synthetic data's baselines, we provide the AE as a baseline for the C-MAPPS dataset. Furthermore, we provide the VAE based on the data $(x_t, \ldots, x_{t-3})$. For the VAEs and the AE, we search for the best latent dimension, number of hidden layers, and learning rate in a hyperparameter search. In addition, we search for the number of latent samples for the VAEs.

We report the Area Under the Receiver Operating Characteristic Curve (AUROC), the AP, and the TPR@{0.05,0.1,0.2,0.3} in Table 4. For the VAEs, the causal approach with CVAEs, and the causal approach consisting of CVAEs and CKDEs, we report the mean and the standard error (SE) of the mean of the respective metrics obtained when retraining the best model configuration ten times.

Our results indicate that the VAE outperforms the KDE when considering the joint probability of all variables concerning all reported metrics. Thus, for the VAEs, we encounter superior performance when considering individual observations instead of a row of four observations. Additionally, the SEs are higher when using multiple observations in a row. Moreover, when considering all variables of the timeseries, the AE outperforms the KDE. In contrast, the VAEs show stronger performance than the AE. In the following, we refer with VAE to the variant relying on one observation. The outperformance becomes especially clear when considering the AP, where the VAE achieves with 0.911 a 14 % higher AP than the AE with 0.798. When we look at the results of the causal approach, we encounter worse results for the variant solely relying on the CVAE for the estimation of the conditional distributions than the baselines AE and VAE. For the AP, the causal variant with CVAEs reaches 0.691 with an SE of ca. 0.09, whereas the VAE achieves 0.911 with an SE of approximately 0.004. The VAE reaches for TPRs@$\{0.05, 0.1\}$ with ca. 0.06 as its highest SE, while for the other metrics, the SE is limited to roughly 0.02. The SE of the causal approach using CVAEs is for all metrics between 0.08 and 0.1, which indicates that the CVAE's ability to estimate the distributions can vary strongly. Thus, we develop an ensemble of CVAEs if the conditional dimension is larger than three and CKDEs otherwise. This variant of the causal approach outperforms the causal variants solely relying on CVAEs or CKDEs. In addition, it outperforms the VAE (one observation) in all metrics by approximately 0.01, by 9.3 % regarding the TPR@0.05, and by 10.9 % regarding the TPR@0.1. In addition, the SE of the causal approach with a mixture of CKDEs and CVAEs is approximately 0.

6 Conclusions

In this paper, we leverage causal information to detect anomalies in multivariate timeseries. Therefore, we develop previous work further by discovering the causal relations of a multivariate timeseries using the CD algorithms adapted to the timeseries setting. Next, the causal relations provide the decomposition of the problem into estimating the conditional distribution given its parents for every variable.

Our results from experiments conducted with synthetically generated data indicate that the causal approach performs better than the VAE, relying on the joint probability of the individual observations. In addition, evaluating graph structures that deviate from the ground truth graph structure indicates the importance of detecting the true causal parents, even when detecting some additional false parents. Furthermore, missing out on the true causal parents leads to strong performance degradation. Additionally, we observe that the performance is less dependent on the causal structure for measurement anomalies than for effect and intervention anomalies.

Table 4 The results for AD on the C-MAPPS dataset. The best results are denoted in bold.

Model	AP	AUROC	TPR@0.05	TPR@0.1	TPR@0.2	TPR@0.3
AE with x_t	0.798	0.926	0.720	0.814	0.878	0.917
VAE with x_t	0.911 ± 0.004	0.976 ± 0.011	0.876 ± 0.060	0.888 ± 0.058	0.969 ± 0.015	0.987 ± 0.008
VAE with $(x_t, \dots, x_{t-3})$	0.859 ± 0.068	0.960 ± 0.023	0.822 ± 0.106	0.867 ± 0.085	0.936 ± 0.046	0.966 ± 0.031
KDE with x_t	0.570	0.711	0.518	0.522	0.542	0.546
Causal + CVAE	0.691 ± 0.086	0.852 ±0.057	0.662 ± 0.096	0.692 ± 0.093	0.754 ± 0.091	0.796 ± 0.082
Causal + CVAE & CKDE	**0.924** **± 0.001**	**0.988** **±0.000**	**0.958** **± 0.001**	**0.985** **± 0.002**	**0.994** **± 0.002**	**0.998** **± 0.000**
Causal + CKDE	0.711	0.955	0.729	0.859	0.963	0.995

Moreover, the experiments on the C-MAPPS dataset show that the causal approach, consisting of an ensemble of CKDEs and CVAEs, performs better than the VAE baselines, which rely on individual observations and rows of observations. We recognize the importance of the individual conditional density estimates being well-fitted since the causal approach, when relying on either one CVAE or CKDEs, performs worse than the mixture.

References

1 Simon D Duque Anton, Sapna Sinha, and Hans Dieter Schotten. Anomaly-based intrusion detection in industrial data with svm and random forests. In *2019 International conference on software, telecommunications and computer networks (SoftCOM)*, pages 1–6. IEEE, 2019.

2 Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE international conference on computer vision*, pages 2745–2754, 2017.

3 Arpita Bhargava and AS Raghuvanshi. Anomaly detection in wireless sensor networks using s-transform in combination with svm. In *2013 5th International Conference and Computational Intelligence and Communication Networks*, pages 111–116. IEEE, 2013.

4 Enyan Dai and Jie Chen. Graph-augmented normalizing flows for anomaly detection of multiple time series. *arXiv preprint arXiv:2202.07857*, 2022. arXiv:2202.07857.

5 Zahra Zamanzadeh Darban, Geoffrey I Webb, Shirui Pan, Charu C Aggarwal, and Mahsa Salehi. Deep learning for time series anomaly detection: A survey. *arXiv preprint arXiv:2211.05244*, 2022. doi:10.48550/arXiv.2211.05244.

6 Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35(5), pages 4027–4035, 2021. doi:10.1609/AAAI.V35I5.16523.

7 Dean K Frederick, Jonathan A DeCastro, and Jonathan S Litt. User's guide for the commercial modular aero-propulsion system simulation (c-mapss). Technical report, NASA, 2007.

8 Felix O Heimes. Recurrent neural networks for remaining useful life estimation. In *2008 international conference on prognostics and health management*, pages 1–6. IEEE, 2008.

9 Jakub Jakubowski, Przemysław Stanisz, Szymon Bobek, and Grzegorz J Nalepa. Anomaly detection in asset degradation process using variational autoencoder and explanations. *Sensors*, 22(1):291, 2021. doi:10.3390/S22010291.

10 Andrew KS Jardine, Daming Lin, and Dragan Banjevic. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical systems and signal processing*, 20(7):1483–1510, 2006.

11 Mohammed Shaker Kareem and Lamia AbedNoor Muhammed. Anomaly detection in streaming data using isolation forest. In *2024 Seventh International Women in Data Science Conference at Prince Sultan University (WiDS PSU)*, pages 223–228. IEEE, 2024.

12 Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

13 Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

14 Daehyung Park, Yuuna Hoshi, and Charles C Kemp. A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder. *IEEE Robotics and Automation Letters*, 3(3):1544–1551, 2018. `doi:10.1109/LRA.2018.2801475`.

15 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. `doi:10.5555/1953048.2078195`.

16 Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

17 Jeffrey S Racine et al. Nonparametric econometrics: A primer. *Foundations and Trends® in Econometrics*, 3(1):1–88, 2008.

18 Ajay Singh Raghuvanshi, Rajeev Tripathi, and Sudarshan Tiwari. Machine learning approach for anomaly detection in wireless sensor data. *International Journal of Advances in Engineering & Technology*, 1(4):47, 2011.

19 Kishore K Reddy, Soumalya Sarkar, Vivek Venugopalan, and Michael Giering. Anomaly detection and fault disambiguation in large flight data: A multi-modal deep auto-encoder approach. In *Annual conference of the phm society*, volume 8(1), 2016.

20 Ferdinand Rewicki, Joachim Denzler, and Julia Niebling. Is it worth it? comparing six deep and classical methods for unsupervised anomaly detection in time series. *Applied Sciences*, 13(3):1778, 2023.

21 Jonas Rothfuss, Fabio Ferreira, Simon Walther, and Maxim Ulrich. Conditional density estimation with neural networks: Best practices and benchmarks. *arXiv preprint arXiv:1903.00954*, 2019.

22 Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1388–1397. Pmlr, 2020. URL: `http://proceedings.mlr.press/v124/runge20a.html`.

23 Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):2553, 2019.

24 Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4(7):487–505, 2023.

25 Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11):eaau4996, 2019.

26 Abhinav Saxena, Kai Goebel, Don Simon, and Neil Eklund. Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 international conference on prognostics and health management*, pages 1–9. IEEE, 2008.

27 Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

28 Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

29 Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search.* MIT press, 2001.

30 Jun-ichi Takeuchi and Kenji Yamanishi. A unifying framework for detecting outliers and change points from time series. *IEEE transactions on Knowledge and Data Engineering*, 18(4):482–492, 2006. `doi:10.1109/TKDE.2006.1599387`.

31 Haowen Xu, Wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Ying Liu, Youjian Zhao, Dan Pei, Yang Feng, et al. Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. In *Proceedings of the 2018 world wide web conference*, pages 187–196, 2018.

32 Asrul H Yaacob, Ian KT Tan, Su Fong Chien, and Hon Khi Tan. Arima based network anomaly detection. In *2010 Second International Conference on Communication Software and Networks*, pages 205–209. IEEE, 2010.

33 Wenzhuo Yang, Kun Zhang, and Steven CH Hoi. A causal approach to detecting multivariate time-series anomalies and root causes. *arXiv preprint arXiv:2206.15033*, 2022.

34 Yufeng Zhang, Jialu Pan, Li Ken Li, Wanwei Liu, Zhenbang Chen, Xinwang Liu, and Ji Wang. On the properties of kullback-leibler divergence between multivariate gaussian distributions. *Advances in Neural Information Processing Systems*, 36, 2024.

35 Hang Zhao, Yujing Wang, Juanyong Duan, Congrui Huang, Defu Cao, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, and Qi Zhang. Multivariate time-series anomaly detection via graph attention network. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 841–850. IEEE, 2020. `doi:10.1109/ICDM50108.2020.00093`.

36 Rui Zhao, Ruqiang Yan, Zhenghua Chen, Kezhi Mao, Peng Wang, and Robert X Gao. Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*, 115:213–237, 2019.

37 Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*, 2018.